Optical Character Recognition

J. BRAUNBECK

What Is Character Recognition?

Before a process can be automated it must be clearly defined. Many of the advantages gained by automation are the result of this need for definition. When asking the question "what is character recognition?" we can again distinguish between "what is recognition?" and "what is character?".

What is Recognition?

Recognition implies the gradation of different signals into classes. A musician is able to classify a tone he hears according to its frequency e. g. as class "c". This class contains violin, piano and trombone tones, low c and top c. These tones are all classed as "c" on the basis of a certain characteristic. Criterion for classification is that the fundamental frequency of these tones is in a known numerical ratio to the reference frequency.

A difficult classification problem is presented by the class "Face of Miss Gerda Schulz". This class comprises Gerda Schulz wearing different hairstyles, hair colours, with and without make-up, as well as black-and-white and coloured photographs. In spite of these variations, the face of Gerda Schulz will always be recognized if one knows her, i.e. if the viewer has memorized the characteristics of the class. This example illustrates another condition important to recognition, namely the necessity of memorizing the specific characteristics of the class which is to be recognized. Colloquially speaking, only things known can be recognized. The extent of memorizing characteristics is even closer defined in colloquial language. If many characteristics have been memorized of the class which is to be recognized, the object is known well.

Somebody who knows Gerda Schulz well will also recognize her even if greatly changed in appearance, e.g. wearing glasses. On the other hand, somebody who knows her less well will either associate her when she wears glasses with another class by mistake, or else not be able to classify her at all. This example illustrates the two possible mistakes which can be made in the recognition process. It is possible to associate an object which is to be recognized with the wrong class, e.g. to mistake a photograph of Gerda Schulz for one of Frieda Maier. In technical language this mistake is termed "substitution".

[Display Engineers are primarily concerned with the display of information to a human rather than to a machine. Mr. Braunbeck's article will be of special interest to those who are concerned with the nature of information and the meaning of the term "recognition" whether such recognition is by man or machine. — Ed.]

The second possibility is the assumption of not knowing the girl in the photograph. Technically speaking this is known as "rejection". Summing up, we can define recognition as the classification of a signal into a definite class on the basis of memorized characteristics. Whether recognition be by man or machine is basically irrelevant for the time being.

What Is a Character?

A character is a signal assigned a definite meaning agreed upon at some time or other. This signal can take any form, consisting, for instance, of a series of electric pulses, a smoke signal or a knot in a string. Character recognition by machine as practiced today is limited to alphanumerical characters. An alphanumerical character is a two-dimensional contrast distributed on a carrier, usually paper, and assigned a definite meaning. Our alphabet, for example, has been assigned sounds from which we form our words. This need not necessarily be the case. Chinese writing, for instance, is constructed so that each symbol corresponds to a term. This has the disadvantage of there being as many groups of characters as there are terms. But it also has the advantage of being independent of the language. If one is familiar with Chinese writing, one is capable of reading it without understanding a word of Chinese.

It is interesting to note that in some instances, such as the operating instructions of machines, the use of symbols has again been reverted to. Our alphabet which goes back to the Phoenicians has also evolved from symbols. "A", for example, is a stylized cattle head. The most important characters in machine reading at present are figures. A figure is a two-dimensional black-and-white distribution, assigned a definite number.

Figure 1 is an example of numbers belonging to the classes "three", "five", and "six" and "eight" together with their intermediate stages. The more liberal the definition of the individual classes, the wider the rounded areas in the figure. The number of rejects in between the classes thus diminishes. At the same time there is a greater risk of assigning a character a wrong class, i.e. obtaining substitutes. This alternative of obtaining many rejects combined with few substitutes on the one hand, or few rejects and many substitutes on the other, is characteristic of the entire recognition technique.

Standardized and Stylized Fonts

Now that we have defined the problem, we can consider how best to impose the task of character recognition on a machine. It should be pointed out from the start that the machine is hardly capable of matching the flexibility of the human brain. It cannot be denied that man is extremely versatile in his ability. He tires easily, however, and is unreliable and slow. The machine on the other hand is more restricted with regard to change, but not susceptible to fatigue or diversion. Furthermore, machines usually work much faster than man doing the same task. It goes without saying that a car is hardly expected to climb stairs, yet assumed to excel by far on the road in speed and perseverance as compared to a pedestrian.

The fact that the machine is less flexible than the human brain is taken into consideration in that characters are standardized and stylized. Standardization largely implies limiting the differences between the characters of a class. Stylization means selecting character form in consideration of machine recognizability.

Figure 2 illustrates type fonts which have been stylized and standardized in three different ways. At the top is the standard

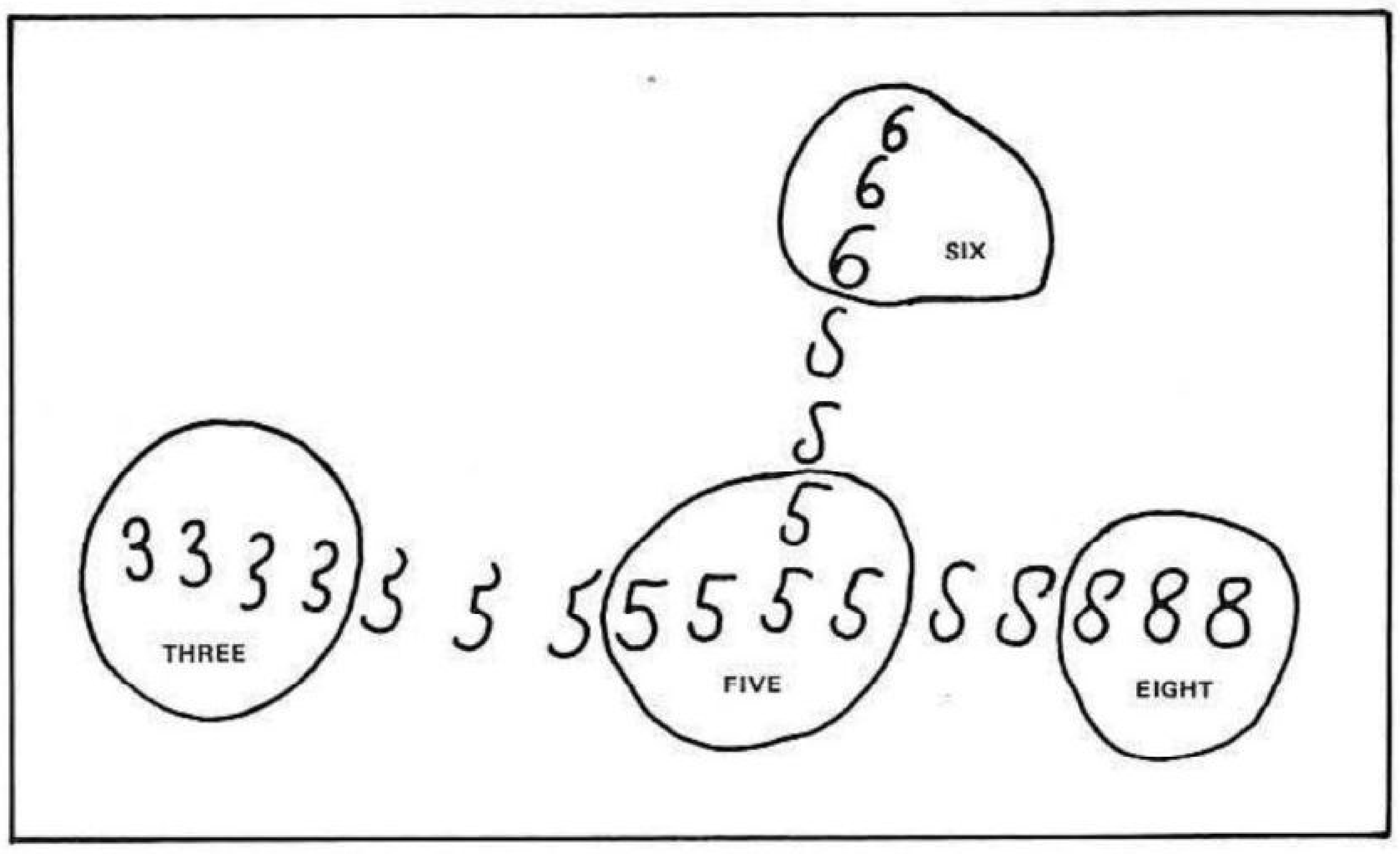

Figure 1: Continuous transition between classes of characters.

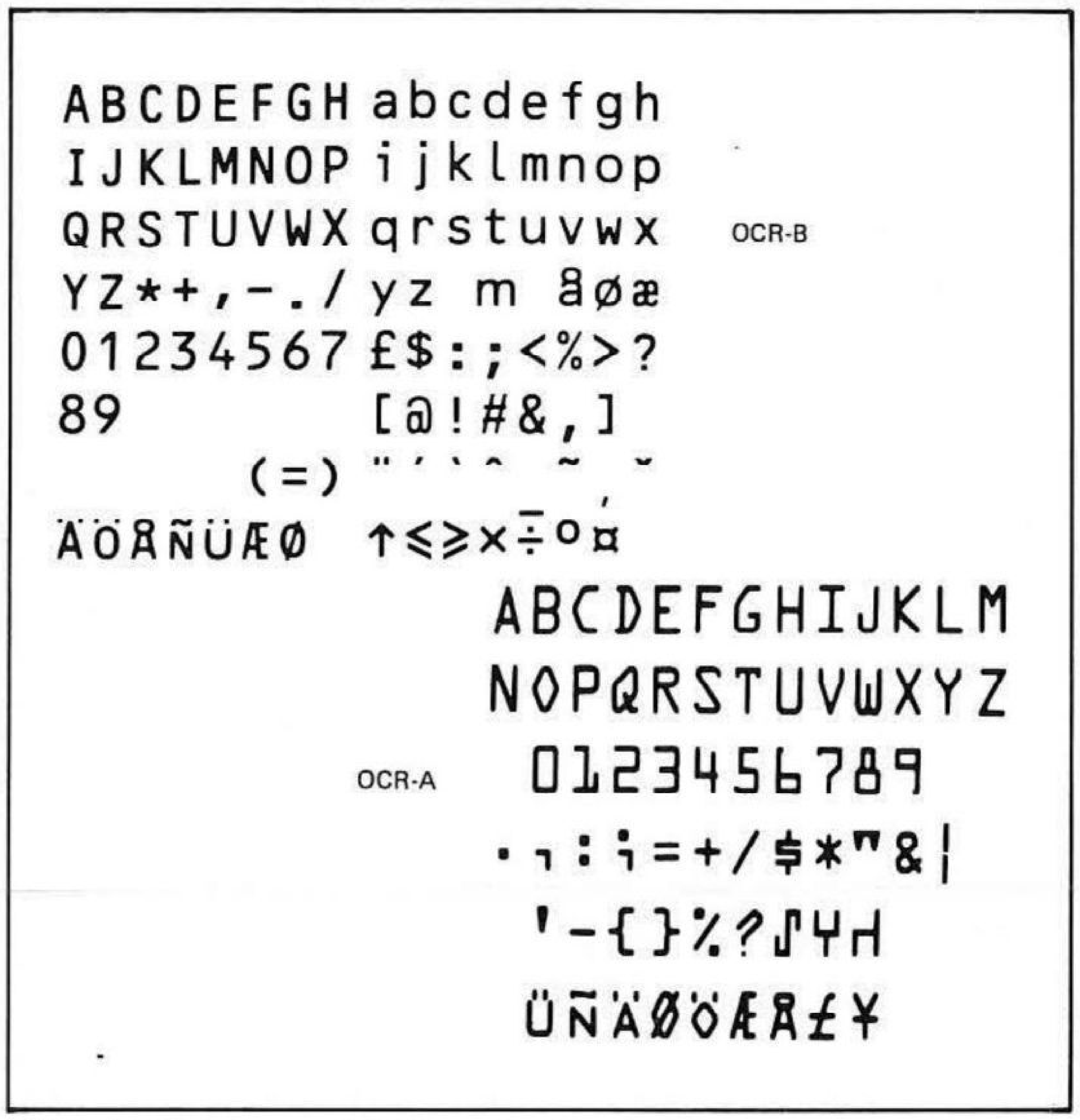

Figure 2: Machine readable type fonts.

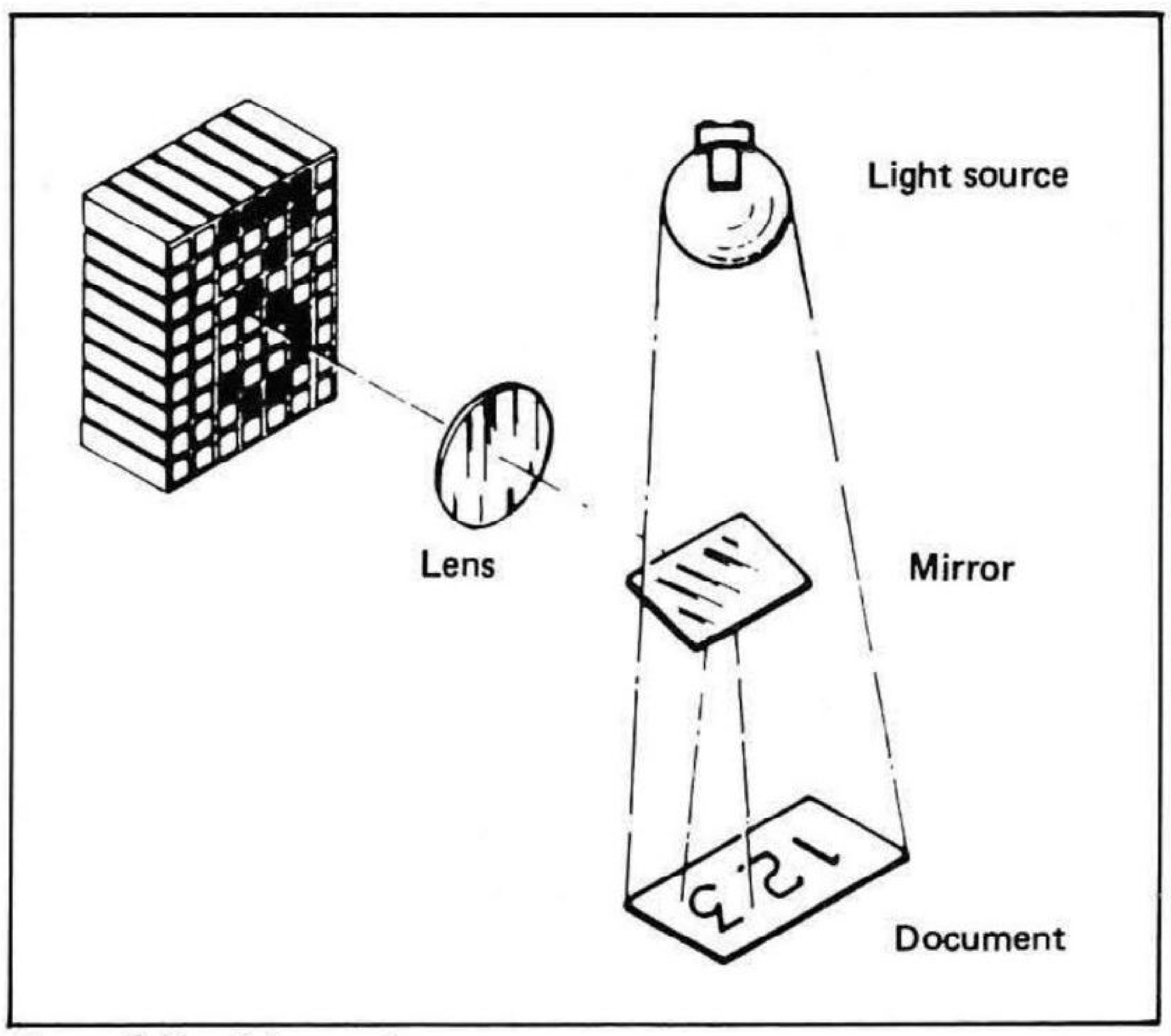

Figure 3: Parallel scanning.

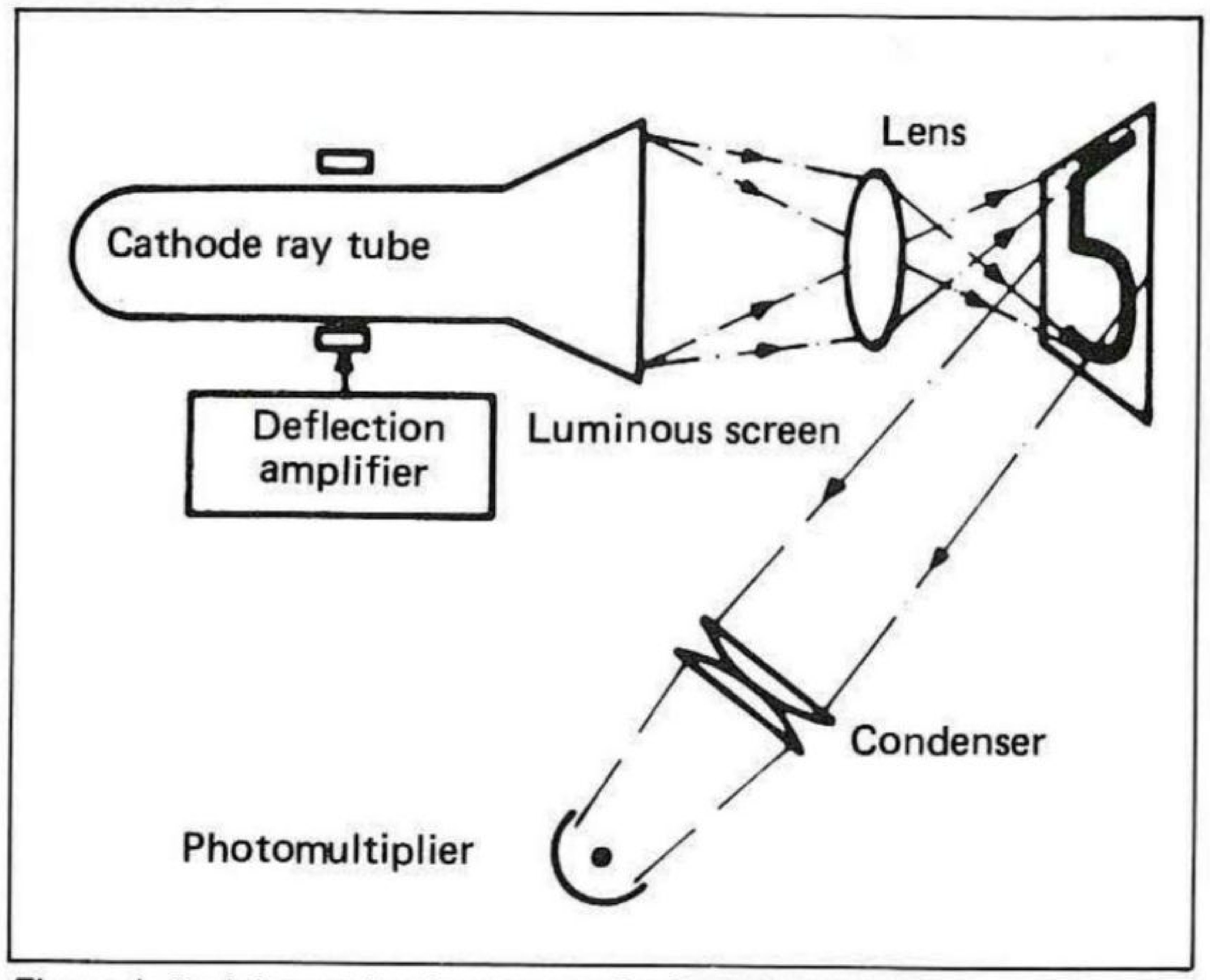

Figure 4: Serial scanning by means of a flying spot tube.

font OCR-B. This font has been only slightly stylized, so that it hardly differs from the usual typewriter or printed characters. The standard font OCR-A appearing below OCR-B is more strongly stylized, facilitating machine reading. This reduction in machine complexity and expenditure has the disadvantage of the characters being somewhat conspicuous. As they seem rather unconventional in appearance at first, their use has been objected to in some fields.

Machine Reading

Automatic reading of characters requires a series of machine parts which are found in some form or other in all character reading devices. The carrier containing the characters to be read must somehow be conveyed to the reading device and be removed again after reading. This necessitates a paper feed assembly. In the devices now in use, very often the paper feed assembly is not part of the reading machine, whereas the reader constitutes a component of the paper feeding device, e.g., the document sorting machine. The characters on paper must then be converted into electrical signals. There is no specific reason for this conversion except that electrical signals can be conveniently worked with in modern technology. This signal conversion is termed scanning. There are various scanning possibilities. Working on the principle of the human eye, each scanning element can be assigned an electric channel. This method, known as "fully parallel scanning", is illustrated in Figure 3. The advantage of fully parallel scanning is the relatively small demand made on the transmission capacity of the individual channels. The information content of the character is distributed over a great number of channels. This is the reason why the eye of the vertebrate is also equipped with a fully parallel scanning system.

If the individual scanning elements are scanned in succession, one transmission channel is all that is required. However, as this channel must process all the information, a high transmission capacity is required of the serial scanning procedure illustrated in Figure 4. Television works on the basis of fully serial scanning as it would hardly be expedient nowadays to operate with a great number of wireless transmission channels in parallel.

Technical solutions are usually found by compromise. The semi-parallel scanning procedure illustrated in Figure 5 is also an example of successful compromise. The character carrier which must in any case be moved for the purpose of transport and removal, horizontally passes along a vertical line of photo cells. Each individual channel need only possess a fraction of the transmission capacity which would be required in serial scanning. On the other hand there is none of the rather considerable line complexity involved in fully parallel scanning.

Modern data processing frequently makes use of the advantages offered by the digital technique. Character readers are no exception. In most of the devices scanning is thus followed by quantization of signals. Halftones, which would in any case play only a subordinate role in recognition, are therefore purposely dispensed with. This renunciation is compensated for by the advantages of the digital technique.

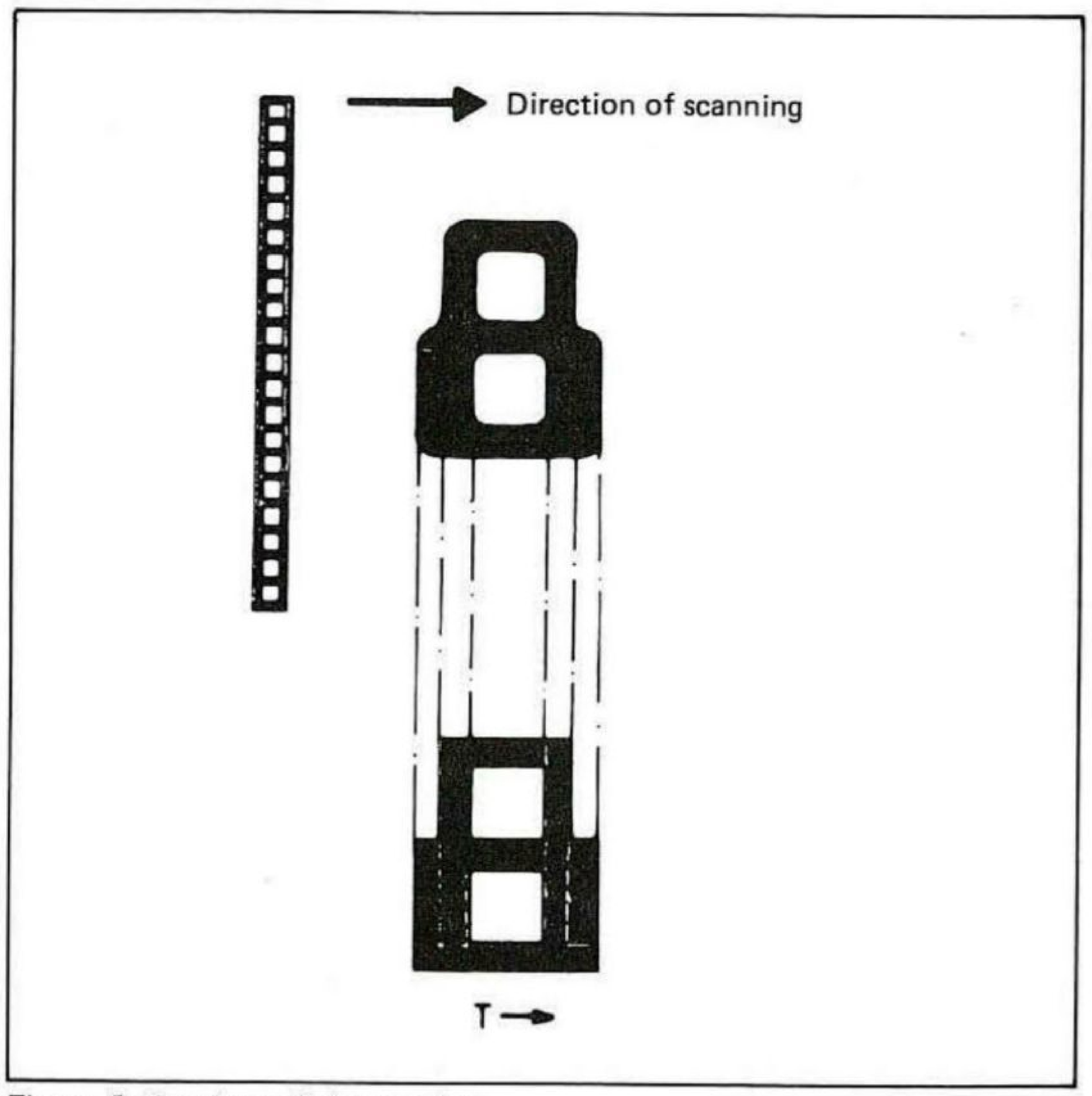

Figure 5: Semi-parallel scanning.

Character Recognition

Character recognition proper, i.e., assigning character classes to the scanned characters, is by means of comparing the electrical signals with samples stored in the device. This comparison can be carried out in a variety of ways, their difference lying in the design complexity involved and the extent of character stylization required.

Stroke analysis, suitable only for strongly stylized characters, such as the numbers of the standard font OCR-A, involves relatively little expenditure. As shown in Figure 6, the scanned character is examined as to vertical and horizontal strokes. In case of semi-parallel scanning, examination is conducted by supervising adjacent channels for dark-spot signals occurring simultaneously. A dark-spot signal occurring in several adjacent channels implies that a vertical stroke has appeared. Each channel is further equipped with a chronometer. A dark-spot signal delivered by a channel within a certain minimum period of time implies that a horizontal stroke has passed this channel. The supervisory circuits described above produce an electrical character description as illustrated in Figure 6. This description is compared with the descriptions of the individual character classes which were stored in the form of diode networks. If it matches one of the descriptions, the recognized character is output in the code used by the machine connected

Continued on page 31

BRAUNBECK—Continued from page 20

to the reader. If the description matches none of the stored descriptions, the reader will transmit a signal of non-recognition, and the character is rejected.

Complex characters, such as those of standard font OCR-B, cannot be represented by horizontal or vertical strokes alone. For recognition by machine the entire character image must be processed. The signals, corresponding to the character and gained through serial, semi-parallel or fully parallel scanning, are used in assembling this character into an "electric image". Imagine a rectangular area fitted with switches instead of tiles. Each one of these light-sensitive electronic switches is now turned on or off depending on whether its associated picture element is white or black. The ideal form of a definite character class which is to be recognized will result in certain selected switches to be turned on. The floating ends of all appertaining resistors are connected to a bus. A "resistor matrix", this being the technical term, of this kind is built up not only for one single character class to be recognized but for all of them. From the fundamental laws of electrical engineering it follows that the resistor matrix in which most of the resistors were switched on shows the highest voltage between its terminals.

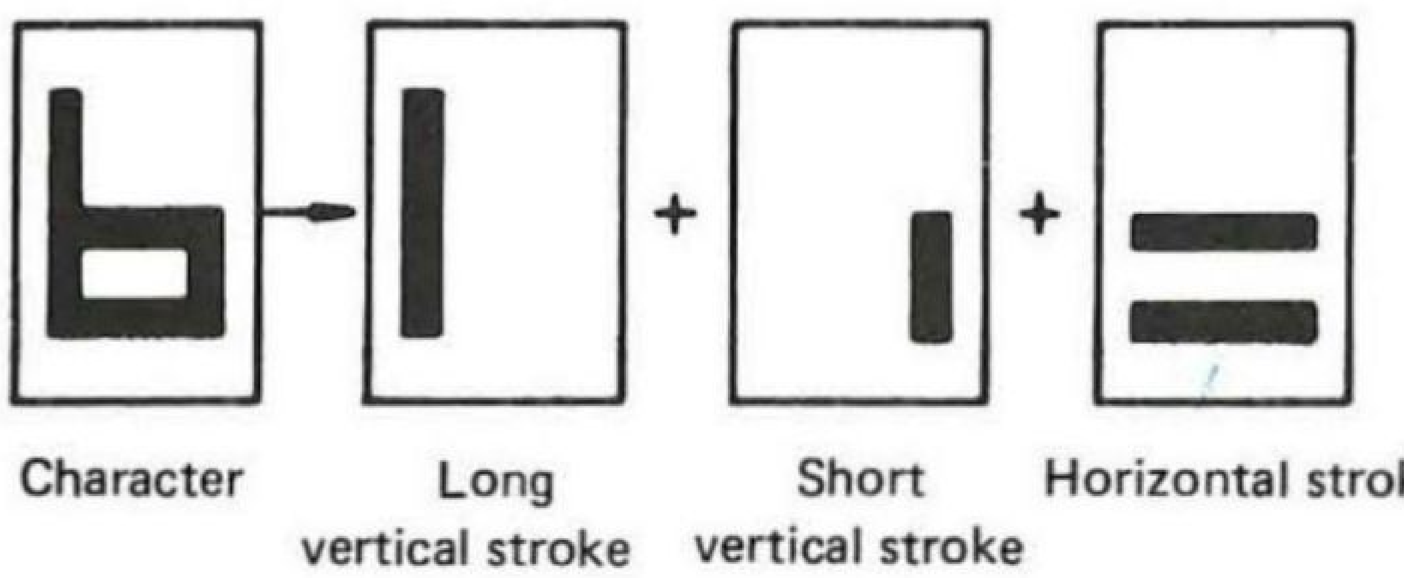

Figure 6: Stroke analysis.

This is verified by connecting the resistor matrices to a so-called maximum filter. The character corresponding to the matrix with the highest voltage is routed by the reader to the follow-up device. If several resistor matrices all yield more or less equally strong voltages, the character is rejected.

The matrix reader can read characters of any shape. There are no specific requirements with regard to stylization except that character pairs may not be too much alike. However, the matrix reader can only read one specific kind of type font. A reader capable of reading a large number of type fonts, technically known as a multifont reader, usually operates on another principle.

A form element reader is a further stage in the development of stroke analysis. The character to be recognized is examined as to the occurrence of form elements. In addition to horizontal and vertical strokes there are curvatures, hooks, open arcs, closed arcs, indentations and other form elements. The form element reader is the most versatile but also the most complex and expensive of all reading machines.

In summing up it may be said that the answer to the question of whether a definite font can be read by machine is almost always in the affirmative today. The question concerning us now is one of design complexity involved in reading a definite font. ■

Dr. Joseph Braunbeck, formerly of Hochhaus, West Germany, is now working on a highly specialized computer system in Vienna, Austria.